"A NEW TOOL DESIGNING FOR CONTENT FILTRATION OF E-LEARNING SITES USING WEB MINING TECHNIQUE "

SYNOPSIS

For the Degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

=====Submitted By======

SANGITA SHIVSHARNAPPA MODI

Under the guidance of

DR. SUDHIR B. JAGTAP

PRINCIPAL

SWAMI VIVEKANAND MAHAVIDYALAYA

UDGIR

SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY, NANDED (M.S.)

Table of Contents

Sr. No	Contents	Page No.
1	Abstract	3
2	Introduction	4
3	Review of Literature	6
4	Objectives of Research	8
5	Methodology	8
6	Significance of Study	9
7	References	9

1. ABSTRACT

Today World Wide Web is the most widely used technology by e-learners. This media is used for collecting, sharing and distributing information. Education is one of the fields where web-based technology has been quickly and successfully adopted. Web mining techniques involves transformation and interpretation of web data to find out the hidden information by using various data mining techniques which will be useful in enhancing E learning environment. In e-learning process learners are accessing different websites for information which is available in images, text, or video format. With this educational content there are some irrelevant data which also displayed on web sites. To block this type of data, effective content-filtering systems are essential. This research is to find the efficient filtering technique, by which learners can use e-learning web sites without any unspecified content.

Keywords: content filter, web mining, e-learning

2. INTRODUCTION

A. Web mining:

Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web. Web mining is the application of data mining technique. Web mining is used to capture relevant information, creating new knowledge out of the relevant data. Web Mining uses data mining techniques to automatically discover and extract information from World Wide Web.

Web content is in textual format which can contained in forum, blogs, mail, social networking, or in ecommerce sites. The image format data content is present in .gif, jpeg, or bmp. While YouTube like application contents are in video formats. Some web content data is in animated format on the web sites which are used for advertising purpose. Web mining is further classified into three:

1. Web content mining
2. Web Structure mining
3. Web Usage mining.

1. Web content mining

Using the objects like text, pictures, multimedia etc. content mining is done in the web. Content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. It is the discovery of useful information from the web documents. It is also distinguishes personal home pages with other web pages. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages.

2. Web Structure mining

Web structure mining, mining is done based on the structure like hyperlinks. This kind of mining emphasizes on the data which describes the structure of the content. It is classified into two types namely intra-page structure and inter-page structure. Intra-page structure means the existence of links within a page. No separate page will be opened in this case. Inter-page structure involves the connection of one page with the other page.

3. Web usage mining

In the case of web usage mining, mining is done on web logs which contain the navigational pattern of users. It focuses on various data mining techniques to understand and analyze search patterns.

B. Web Content filtering

A content filter can be defined as a program that helps system administrators decide which content is acceptable for viewing and accessing. With the implementation of this program, system administrators can prevent people accessing or viewing certain items on the web. These items could be harmful if opened or accessed. The most common items that are content filtered include, executable files and websites. The most general example of content

filtering is blocking the websites containing pornographic material or the social networking websites which are not related to work.

C. Web Usage Mining Technique

Four main Web usage mining techniques are shown here: association rules, sequential patterns, supervised classification and clustering.

1. **Association Rules**:- Association rules are if/then statements that help to uncover relationship between seemingly unrelated data in a relational database or other information repository. The target is to find pages that are accessed together by majority of the user and hence should be linked in a proper way in order to maximize user satisfaction by providing to the user the access flow they expect.

2. **Sequential Patterns**:- Frequent subsequence among large amount of sequential data is discovered using Sequential patterns. Sequential patterns are employed to find sequential navigation patterns that appear in users' sessions frequently, during Web usage mining process.

3. **Supervised Classification**:- Supervised classification can be used in predicting the class to which an object or individual is likely to belong. Supervised classification is suitable to apply if the data is known to have a small number of classes or group, the classes or groups are already known and some training data with their classes known is available.

4. **Clustering Algorithms** - K-means, Suffix Tree and LINGO

The performance of the web search engines could be improved by properly clustering the search result documents. Most of the users are not capable to give the appropriate query to get what exactly they wanted to retrieve. So the search engine will retrieve a massive list of data, which are ranked by the page rank algorithm or relevancy algorithm or human judgment algorithm. The user will always find to self with the unrelated information related to the search due to the ambiguity in the query by the user. Evaluating the performance of the clustering algorithm is not as trivial as counting the number of errors or the precision and recall of a supervised classification algorithm.

D. Web Content mining tools

With the flood of information and data on the Web, the content mining tools helps to extract the essential information that one would require. Some of them are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda, and Web Content Extractor.[6]

- Screen-scaper : Screen-scraping is a tool for extracting/mining information from web sites. It can be used for searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. Screen-scraper present a graphical interface allowing the user to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data.[7]

- Automation Anywhere 6.1 (AA) : AA is a Web data extraction tool used for retrieving web data, screen scrape from Web pages or use it for Web mining. It is a technique SMART Automation Technology for fast automation of complex tasks.[8]

E. E-learning technology

E-learning offers the ability to share material in all kinds of formats such as videos, slideshows, word documents and PDFs. Conducting webinars (live online classes) and communicating with professors via chat and message forums is also an option available to users. E-learning offers an alternative that is faster, cheaper and potentially better. Because of blogs, forum, social networking, YouTube, educational lots of data uploaded and can make available for students of different faculty. In this easy learning process any one can extracting research papers, articles, journals, e-books, e-library for their study. But while extracting study material irrelevant information like adult sites, videos, advertises of ecommerce sites are displayed on the sites, which are major barriers of e-learning technology. To clean these unwanted data from e-learning web sites content filtration is used using web mining techniques.

3. REVIEW OF LITERATURE

Sagar More[1] 2014, proposed a method for mining path traversal patterns. The modified algorithm use the filtering technique which removes duplicate and unwanted data in web browsing session and show the effectiveness as compared to formal algorithms.

Anshul Bhargav and Munish Bhargav [2] 2014, proposed a framework which is based on three steps: pre-processing, pattern discovery and users classification,that helps the website Administrators to better serve the needs of their websites users.

Jia-Ching, Chu-Yu and Vincent S. [3] 2012, proposed an efficient incremental data mining algorithm named id-Matrix-Miner for mining web navigation patterns with dynamic threshold.

Anuradha Yadav, Satbir Jain [7] 2011, proposed that how by using WUM techniques we can make e-learning environment more effective. Based on WUM technique how we can design an effective e-learning environment and implementation of the model builds more intelligent system and can promote interest of learners or learning environment is modifying its behavior to reflect learning style.

Ling Guo, Xin Xiang, YuanChun Shi [8] 2004, states that online learning activities, also present approach to assess student behavior through web log mining.

Bhavani Thuraisingham[9] 2004 states that Multimedia is more complex-as the sequence of progresses, the concept being may change as well. Understanding and representing changes in the mining process is necessary to mine multimedia data.

P. Y. Lee, Nanyang, S. C. Hui, A. C. M. Fong[10] 2002 focus on blocking pornography because it is among the most prolific and harmful Web content. However, a framework is adaptable for filtering other objectionable Web material.

A. F. Drimbarean,P. M. Corcoran ; M. Cuic ; V. Buzuloiu[8]2001,describes image-processing the techniques include a novel method of matching skin tones based on a fuzzy classification scheme and shape recognition techniques to match faces and other elements of the human anatomy.

Keiichiro Hoashi [9]2000, proposes the use of a non-relevant information profile in order to retrieve more relevant documents without excessive retrieval of non-relevant documents.

Yoshitaka A. and Ichikawa T.[10]1999, disused that multimedia database system includes a multimedia database management system (MMDBMS),which manages and also provides support for storing, manipulating and retrieving multimedia data from a multimedia database, a large collection of multimedia objects, such as image, video, audio and hypertext data.

4. OBJECTIVES OF RESEARCH

The objectives of the study are as following:

1. Searching of e-learning sites which contain the e-books, tutorials, audio notes, video lecturers etc.
2. Searching different types of contents present on e-learning sites.
3. Study of web mining techniques.
4. Web content identification.
5. Data pattern recognisation.
6. Designing a tool for content filtration.
7. Filtering the unwanted pattern using the tool from E-learning websites.

5. METHODOLOGY

Firstly, relevant literature was reviewed, including basic information about the certain study area, knowledge about previous use of the web mining techniques. Secondly, finding the various web sites of e-learning which requires filter. Find the web content types present on the sites with the help of survey method on internet. Study the different present filtration tools and techniques. Build new content filtration tool for e-learning web sites which are used by learners. Implement the filtration algorithm to check the efficiency.

6. SIGNIFICANCE OF THE STUDY

This research improves e-learning concept which motives self learning skills in the students. It reduces the sites hitting. Filtering gives the learner educational contents only, instead of unwanted data. The tool will help the user to surf the e learning sites with more ease and safety, by preventing the non-related and unwanted content which is available on such type of e-learning sites.

7. REFERENCES:

[1] More, S. 2014. Modified Path Traversal for an Efficient Web Navigation Mining. International Conference on Advanced Communication Control and Computing Technologies (pp. 940-945). IEEE. DOI: 10.1109/ICACCCT.2014.7019232

[2] Bhargav, A., & Bhargav, M. 2014. Pattern Discovery and Users Classification Through Web Usage Mining. International Conference on Control, Instrumentation, Communication and Computational Technologies (pp. 632-636).
IEEE. DOI:10.1109/ICCICCT.2014.6993038.

[3] Jia-Ching, Ying, Chu-Yu, Chin & Tseng, Vincent S. 2012. Mining Web Navigation Patterns with Dynamic Thresholds for Navigation Prediction. International Conference on Granular Computing 4-619. IEEE. DOI: 10.1109/GrC.2012.6468696

[4] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi Overview of Web Content Mining Tools Volume 2 , 2013

[5] Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.

[6] Automation Anywhere Manual. AA, http://www.automationanywhere.com Viewed 06 February 2013.

[7] Anuradha Yadav, Satbir Jain, April 2011. “Analyses of web usage mining techniques to enhance the capabilities of E-learning environment” published by IEEE.

[8] Ling Guo, Xin Xiang, YuanChun Shi ,”Use Web Usage Mining to Assist Online E-Learning Assessment” Published in:Advanced Learning Technologies, 2004.
Proceedings IEEE International Conference.

[9] Bhavani Thuraisingham, “Managing and Mining Multimedia Databases” at International Journal on Artificial Intelligence Tools Vol. 13, No. 3 (2004) 739-759

[10] P. Y. Lee ,S. C. Hui, A. C. M. Fongoct2002 “Neural networks for web content filtering” Published in: IEEE Intelligent Systems (Volume:17 , Issue: 5)State Key Lab of Intelligent Technology and Systems, published by IEEE

[11] A. F. Drimbarean, P. M. Corcoran, M. Cuic, V. Buzuloiu “Image processing techniques to detect and filter objectionable images based on skin tone and shape recognition” Published in:Consumer Electronics, 2001. ICCE. International Conference on by IEEE

[12] Keiichiro Hoashi Kazunori Matsumoto Naomi Inoue Kazuo Hashimoto "Document Filtering Method Using Non-Relevant Information" In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp.176-183, 2000.

[13] Yoshitaka A. and Ichikawa T., "A Survey on Content-based Retrieval for Multimedia Databases".IEEE Transaction on Knowledge & Data Engineering, Vol 11, 1999, pp. 81-93

Sangita S. Modi
Research Student

Dr. Sudhir B. Jagtap
Research Guide